

А.А. Россиев

**ИТЕРАЦИОННОЕ МОДЕЛИРОВАНИЕ НЕПОЛНЫХ
ДАННЫХ С ПОМОЩЬЮ МНОГООБРАЗИЙ
МАЛОЙ РАЗМЕРНОСТИ**

Красноярск – 2000

Итерационное моделирование неполных данных с помощью многообразий малой размерности. Россиев А.А.

Описывается метод моделирования неполных данных с помощью последовательности кривых, обобщающий метод главных компонент. Обсуждаются три версии метода:

а) линейный – с моделированием данных последовательностью линейных многообразий малой размерности;

б) квазилинейный – с построением "главных кривых" (или "главных поверхностей"), однозначно проектируемых на линейные главные компоненты;

в) существенно нелинейный – основанный на построении "главных кривых" ("главных струн и балок") с использованием вариационного принципа; итерационная реализация этого метода близка методу самоорганизующихся карт Кохонена.

Все полученные зависимости, кроме линейной, нуждаются в экстраполяции, которая производится с помощью формул Карлемана. Метод трактуется как построение нейросетевого конвейера, решающего следующие задачи:

а) заполнение пробелов в данных;

б) ремонт данных – корректировка значений исходных данных так, чтобы наилучшим образом работали построенные модели;

в) построение вычислителя, заполняющего пробелы в поступающей на вход строке данных (в предположении, что данные о новых объектах связаны теми же самыми отношениями, что и в исходной таблице).

Разработанная технология предназначена для решения широкого спектра задач, связанных с обработкой неполных данных. Она реализована в программных продуктах FAMaster и ModelAnalyzer.

ВВЕДЕНИЕ	5
ГЛАВА I. ПРОБЛЕМЫ ОБРАБОТКИ И ИТЕРАЦИОННОГО МОДЕЛИРОВАНИЯ НЕПОЛНЫХ ДАННЫХ С ПОМОЩЬЮ МНОГООБРАЗИЙ МАЛОЙ РАЗМЕРНОСТИ	15
ВВЕДЕНИЕ.....	15
I.1. МЕТОДЫ ОБРАБОТКИ ДАННЫХ С ПРОПУСКАМИ.....	15
I.2. ГЛАВНЫЕ КРИВЫЕ	17
I.3. ТАБЛИЦЫ ЭМПИРИЧЕСКИХ ДАННЫХ.....	17
<i>Задачи эмпирического предсказания.....</i>	<i>18</i>
<i>Требования к методам обработки таблиц эмпирических данных.....</i>	<i>19</i>
I.4. ИТЕРАЦИОННОЕ МОДЕЛИРОВАНИЕ НЕПОЛНЫХ ДАННЫХ С ПОМОЩЬЮ МНОГООБРАЗИЙ МАЛОЙ РАЗМЕРНОСТИ.....	20
<i>Постановка задачи</i>	<i>22</i>
ГЛАВА II. ЛИНЕЙНЫЕ И КВАЗИЛИНЕЙНЫЕ МНОГООБРАЗИЯ МАЛОЙ РАЗМЕРНОСТИ.....	24
ВВЕДЕНИЕ.....	24
II.1. МЕТОД ГЛАВНЫХ КОМПОНЕНТ	24
II.2. ГЕОМЕТРИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ	25
II.3. ЛИНЕЙНЫЕ МНОГООБРАЗИЯ МАЛОЙ РАЗМЕРНОСТИ.....	26
<i>Сингулярное разложение матриц с пропусками</i>	<i>26</i>
<i>Метод главных компонент для таблиц с пробелами.....</i>	<i>28</i>
II.4. ВОССТАНОВЛЕНИЕ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ЛИНЕЙНЫХ МОДЕЛЕЙ....	31
II.5. МНОГОМЕРНЫЕ ЛИНЕЙНЫЕ МНОГООБРАЗИЯ	32
<i>Ортогонализация базисной системы векторов</i>	<i>32</i>
<i>Двумерные линейные модели.....</i>	<i>33</i>
<i>Трёхмерные линейные модели</i>	<i>34</i>
II.6. КВАЗИЛИНЕЙНЫЕ МНОГООБРАЗИЯ МАЛОЙ РАЗМЕРНОСТИ	35
<i>Метод построения квазилинейных моделей</i>	<i>35</i>
II.7. ИНТЕРПОЛЯЦИЯ	36
<i>Интерполяция полиномом небольшой степени</i>	<i>36</i>
<i>Интерполяция кубическими сплайнами.....</i>	<i>38</i>
II.8. ЭКСТРАПОЛЯЦИЯ	40
<i>Проблема экстраполяции, оптимальное аналитическое продолжение и формула Карлемана.....</i>	<i>40</i>
<i>Интерполяция и оптимальное сглаживание по формуле Карлемана.....</i>	<i>42</i>

II.9. ИСПОЛЬЗОВАНИЕ КВАЗИЛИНЕЙНЫХ МОДЕЛЕЙ.....	43
II.10. МЕХАНИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ.....	43
II.11. НЕЙРОННЫЙ КОНВЕЙЕР ДЛЯ ДАННЫХ С ПРОПУСКАМИ	45
ГЛАВА III. САМООРГАНИЗУЮЩИЕСЯ МНОГООБРАЗИЯ МАЛОЙ	
РАЗМЕРНОСТИ.....	47
ВВЕДЕНИЕ.....	47
III.1. ОПРЕДЕЛЕНИЕ ГЛАВНЫХ КРИВЫХ.....	47
<i>Алгоритм Hastie-Stuetzle</i>	48
III.2. ИДЕЯ САМООРГАНИЗУЮЩИХСЯ КРИВЫХ.....	49
III.3. SOC	50
<i>Алгоритм построения SOC</i>	51
III.4. SOM.....	52
<i>Квадратная SOM</i>	52
<i>Гексагональная SOM</i>	53
<i>Алгоритм построения квадратной SOM</i>	54
<i>Алгоритм построения гексагональной SOM</i>	55
III.5. ПРОБЛЕМА ЛОКАЛЬНОГО МИНИМУМА	56
<i>Метод отжига</i>	56
<i>Многосеточный метод</i>	56
III.6. СГЛАЖИВАНИЕ	58
<i>Проблема сглаживания</i>	58
<i>Кусочно-линейная проекция на ломаную</i>	59
<i>Кусочно-линейная проекция на квадратную и гексагональную сетки</i>	63
III.7. МЕХАНИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ.....	64
ГЛАВА IV. ЭКСПЕРИМЕНТАЛЬНЫЕ РЕЗУЛЬТАТЫ.....	65
ВВЕДЕНИЕ.....	65
IV.1. ФАКТОРНЫЙ И КЛАСТЕРНЫЙ АНАЛИЗ АДМИНИСТРАТИВНЫХ ТЕРРИТОРИЙ КРАСНОЯРСКОГО КРАЯ ПО ПОКАЗАТЕЛЯМ ЗДОРОВЬЯ И ЗДРАВООХРАНЕНИЯ	65
IV.2. ТАБЛИЦА РЕЗУЛЬТАТОВ ВЫБОРОВ ПРЕЗИДЕНТОВ США.....	68
IV.3. ВЕРИФИКАЦИЯ СВЯЗЕЙ МЕЖДУ ДВУМЯ ДИНАМИЧЕСКИМИ СИСТЕМАМИ.....	71
IV.4. СОЧЕТАННЫЕ ПОРАЖЕНИЯ ПРОВОДЯЩЕЙ СИСТЕМЫ СЕРДЦА	74
ЗАКЛЮЧЕНИЕ	76
ЛИТЕРАТУРА	78

ВВЕДЕНИЕ

Практически любое научное исследование связано со сбором и обработкой данных, в большинстве случаев структурируемых в виде таблиц. Достаточно часто таблицы содержат пробелы (пустые ячейки), возникающие вследствие невозможности наблюдения, отсутствия необходимых инструментов и т.п. Кроме пробелов, по тем же причинам, а также из-за неточных входных данных, из-за погрешностей, вносимых на отдельных этапах вычислений, погрешностей самих методов вычислений, данные могут содержать искаженные значения (например, случайный шум).

Большинство известных методов анализа данных не могут обрабатывать такую информацию. Иногда на место пропущенных значений можно попытаться поставить среднее значение данного показателя по всей таблице, либо исключить из таблицы показатель (столбец) или объект (строку) с пробелом. Но и то, и другое приводит к потере информации или к ее значительному искажению, а иногда и вообще делает решение задачи невозможным.

Несмотря на перечисленные трудности, данные нуждаются в обработке. Поэтому явно или неявно возникает необходимость в процедуре заполнения пропусков и ремонта (здесь и далее под "ремонтом" подразумевается поиск и исправление неправдоподобных значений) данных – процедуре предобработки.

Следует особо подчеркнуть, что, рассматривая данные проблемы, невозможно говорить ни об истинных значениях данных, ни даже о статистической доказательности, но только о правдоподобии. Особую трудность (и в то же время особый интерес) описанные задачи имеют в тех случаях, когда плотность пробелов высока, расположены они нерегулярно, а данных немного, например, число строк таблицы примерно равно числу столбцов.

Рассматривая множество данных как набор точек (линейных многообразий при наличии пропусков) в пространстве, можно строить их линейные и нелинейные приближения – модели, с помощью которых можно восстанавливать имеющиеся пробелы и ремонтировать данные.

Кроме этого, такие модели позволяют проводить дополнительные эксперименты, которые по тем или иным причинам невозможны с реальным объектом исследования. В частности, представляет большой интерес задача оценки правдоподобности полученных данных.

В настоящем исследовании описывается метод моделирования неполных данных с помощью последовательности кривых, обобщающий метод главных компонент. Он также может быть интерпретирован как процедура сингулярного разложения матриц с пробелами. Обсуждаются три версии метода, которые имеют ясную физическую интерпретацию:

а) линейный – с моделированием данных последовательностью линейных многообразий малой размерности;

б) квазилинейный – с построением "главных кривых" (или "главных поверхностей"), однозначно проектируемых на линейные главные компоненты;

в) существенно нелинейный – основанный на построении "главных кривых" ("главных струн и балок") с использованием вариационного принципа; итерационная реализация этого метода близка методу самоорганизующихся карт Кохонена.

Все полученные зависимости, кроме линейной, нуждаются в экстраполяции, которая производится с помощью формул Карлемана. Метод трактуется как построение нейросетевого конвейера, решающего следующие задачи:

а) заполнение пробелов в данных;

б) ремонт данных – корректировка значений исходных данных так, чтобы наилучшим образом работали построенные модели;

в) построение вычислителя, заполняющего пробелы в поступающей на вход строке данных (в предположении, что данные о новых объектах связаны теми же самыми отношениями, что и в исходной таблице).

Разработанная технология итерационного моделирования неполных данных с помощью многообразий малой размерности предназначена для решения широкого спектра задач, связанных с обработкой неполных данных. Она реализована в программных продуктах FAMaster и ModelAnalyzer.

В **первой главе** работы дан обзор литературы по существующим методам обработки данных с пробелами, а также по теории главных кривых. В частности, рассматриваются методы регрессионного анализа, методы максимального правдоподобия, эмпирические методы и нейросетевые методы заполнения пропусков в таблицах.

Далее приводятся теоретические и методологические основы моделирования данных с пробелами многообразиями малой размерности. При этом вначале рассматриваются эмпирические таблицы и задачи эмпирического предсказания, требования к методам обработки таблиц эмпирических данных, а затем описываются теоретические основы самого метода моделирования данных с пробелами многообразиями малой размерности.

Пусть задана прямоугольная таблица $A=(a_{ij})$ типа "объект-признак", где строки соответствуют объектам, а столбцы – признакам, и пусть часть информации в таблице отсутствует (т.е. некоторые $a_{ij}=@$, где @ – значок, означающий отсутствие данных).

Тогда каждая строка таблицы – это вектор данных x с k пробелами, который представляется как k -мерное линейное многообразие L_x , параллельное k координатным осям, которые соответствуют пропущенным данным. При наличии априорных ограничений на пропущенные значения место L_x занимает прямоугольный параллелепипед $P_x \subset L_x$. Ищется многообразие M заданной малой размерности (чаще всего – кривая), наилучшим образом приближающее данные и удовлетворяющее некоторым требованиям регулярности. Из данных вычитаются ближайшие к ним точки многообразия M – получается остаток – и процесс повторяется, пока остатки не приблизятся в достаточной степени к

нулю. Дальнейшая конкретизация метода состоит в указании того, как строится многообразие M .

Приводится связь предложенного алгоритма с методами главных кривых и сингулярного разложения таблиц данных, рассматриваются некоторые условия применимости и, наконец, приводится постановка задачи.

Требуется построить модели, которые позволяли бы решать следующие три задачи, связанные с восстановлением пропущенных данных:

- 1) правдоподобно заполнить имеющиеся пробелы в данных;
- 2) отремонтировать данные, т.е. исправить их значения таким образом, чтобы наилучшим образом работали построенные модели;
- 3) построить по имеющейся таблице вычислитель, который бы заполнял пробелы в данных и ремонтировал бы их по мере поступления (в предположении, что данные в поступающей на вход строке связаны теми же соотношениями, что и в исходной таблице).

Вторая глава "Линейные и квазилинейные многообразия малой размерности" посвящена методу моделирования неполных данных с помощью линейных и квазилинейных многообразий малой размерности.

Первоначально дается краткое описание классического метода главных компонент.

Далее рассматривается геометрическая интерпретация метода итерационного моделирования неполных данных с помощью линейных и квазилинейных многообразий малой размерности.

Возьмем за основу прямую $f(x)=xy+b$, которая задается направляющим вектором y и проходит через точку, определяемую вектором b , и расположим ее так, чтобы она наилучшим (в некотором точном смысле) образом приближала исходные данные. Исходные данные однозначно проецируются на полученную прямую, при этом проекция определяется как ближайшая к данному точка прямой (как для полных, так и для неполных данных). Если из данных вычесть их проекции, получим отклонения от первой модели (в предположении, что прямая моделирует данные). Это множество отклонений также можно смоделировать прямой данного вида, спроецировать на нее и вычесть проекции – и так далее. Над полученной прямой можно построить кривую – квазилинейную модель, где в качестве аргумента гладкой вектор-функции выступает значение проекции на исходную прямую. В этом и заключается итерационный процесс моделирования неполных данных линейными и квазилинейными многообразиями малой размерности.

Пусть задана прямоугольная таблица $A=(a_{ij})$, ячейки которой заполнены действительными числами или значком @, означающим отсутствие данных.

Требуется представить исходную матрицу A в виде суммы матриц P_q : $A = \sum_q P_q$, где каждая P_q имеет вид $x_i y_j + b_j$.

Если потребовать, чтобы вектор b всегда был нулевым (то есть, чтобы моделирующая прямая проходила через начало координат), то получим сингулярное разложение матриц с пропусками, иначе имеем метод главных компонент для неполных данных.

Основная процедура – поиск наилучшего приближения таблицы A , содержащей пропуски, матрицей вида $x_i y_j + b_j$ методом наименьших квадратов:

$$\Phi = \sum_{\substack{i,j \\ a_{ij} \neq @}} (a_{ij} - x_i y_j - b_j)^2 \rightarrow \min. \quad (1)$$

Решение дается последовательными итерациями по явным формулам – при фиксированных значениях одних переменных находятся значения других. Этот итерационный процесс является сходящимся – так как на каждом этапе значение функционала уменьшается.

Таким образом, для данной матрицы A находим наилучшее приближение матрицей P_1 вида $x_i y_j + b_j$. Далее, из матрицы A вычитаем полученную матрицу P_1 , и для полученной матрицы уклонений $A - P_1$ вновь ищем наилучшее приближение P_2 этого же вида и т.д. Контроль ведется, например, по остаточной дисперсии столбцов.

В результате исходная матрица данных A представляется в виде суммы матриц P_q , т.е. $A = P_1 + P_2 + \dots + P_q$.

С использованием Q полученных факторов (матриц заданного вида) можно решать задачи заполнения пропусков в таблице и ремонта искаженных значений:

Q-факторное заполнение пропусков: пропущенные значения в исходной матрице A определяются из суммы Q полученных матриц вида $x_i y_j + b_j$;

Q-факторный "ремонт" таблицы: значения в исходной матрице заменяются на сумму Q полученных матриц вида $x_i y_j + b_j$.

В функционале (1) используется матрица данных, но если переписать его для вектора данных, то получается процедура заполнения пропусков и ремонта в отдельно взятом векторе.

Совершенно аналогично строятся линейные многообразия большей размерности – два и три. Однако им свойственны некоторые особенности – например, образующие их векторы не обязательно ортогональны и может потребоваться процедура их ортогонализации, а также может возникнуть необходимость в процедуре вращения – для более наглядного представления.

Метод построения квазилинейных моделей основывается на методе построения соответствующих линейных моделей. Его предлагается проводить в три этапа.

1. Построение линейной модели.

2. Интерполяция (сглаживание): строится вектор-функция $f(t)$, минимизирующая функционал:

$$\Phi = \sum_{\substack{i,j \\ a_{ij} \neq @}} (a_{ij} - f_j(\sum_k a_{ik} y_k))^2 + \alpha \int_{-\infty}^{+\infty} (f''(t))^2 dt,$$

где $\alpha > 0$ – параметр сглаживания.

3. Экстраполяция. Решение этой задачи возможно провести несколькими способами, но наиболее интересным представляется использование такой экстраполяции, где сглаженная вектор-функция $f(t)$ экстраполируется с некоторого конечного множества $\{t_k\}$ (которое не обязательно связано с

проекциями на прямую $z_j=ty_j+b_j$ исходных строк данных) на всю вещественную прямую с использованием формул Карлемана (с помощью формул Карлемана экстраполируется отклонение кривой $f(t)$ от прямой $ty+b$):

$$f(t) \approx ty + b + \sum_{k=1}^m (f(t_k) - t_k y - b) \frac{2(e^{\lambda t} - e^{\lambda t_k})}{\lambda(e^{\lambda t} + e^{\lambda t_k})(t - t_k)} \prod_{\substack{j=1 \\ j \neq k}}^m \frac{(e^{\lambda t_k} + e^{\lambda t_j})(e^{\lambda t} - e^{\lambda t_j})}{(e^{\lambda t_k} - e^{\lambda t_j})(e^{\lambda t} + e^{\lambda t_j})}, \quad (2)$$

где λ – параметр метода, характеризующий, насколько широка полоса на плоскости комплексных чисел, в которой гарантированно голоморфна экстраполируемая функция (эта ширина равна π/λ).

Приводятся различные решения задачи интерполяции: с помощью полиномов небольшой степени и кубических сплайнов.

Далее подробно рассматриваются проблемы экстраполяции, оптимального аналитического продолжения. В качестве решения этих проблем предлагается использовать формулу Карлемана, которую можно применять не только для экстраполяции, но и для интерполяции.

Процедура использования квазилинейных моделей несколько отличается от аналогичной процедуры в линейном случае.

Точка на построенной кривой $f(t)$, соответствующая полному ("комплектному") вектору данных a , строится как $f((a, y))$. В этом и заключается квазилинейность метода: сначала ищется проекция вектора данных на прямую $Pr(a)=ty+b$, $t=(a, y)$, а затем строится точка на кривой $f(t)$. Также и для неполных векторов данных – сначала ищется на прямой ближайшая к ним точка $t(a)$ (проекция неполного вектора a), а затем – соответствующая точка на кривой $f(t)$ при $t=t(a)$. Следующим шагом из данных вычитаются их проекции на построенную кривую, то есть матрица данных заменяется на матрицу уклонений от модели. Далее процедура продолжается до тех пор, пока не выполнится некоторое условие остановки (например, близость уклонений к началу координат).

В результате исходная таблица предстает в виде Q -факторной модели:

$$a_{ij} \cong \sum_q f_j^q(t_i^q). \quad (3)$$

Описанные методы построения многообразий малой размерности, наилучшим образом приближающих данные, имеют ясные механические представления, лежащие в их основе.

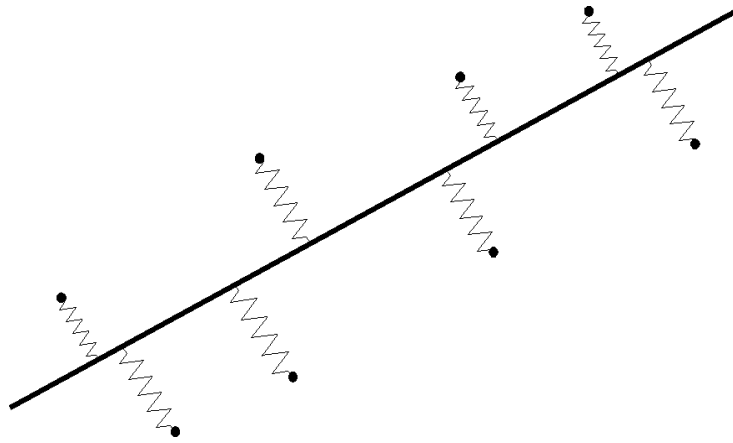


Рис. 1.

В линейном случае, в пространство данных помещена прямая жесткая балка (рис. 1), свободно соединенная пружинками с данными. Минимум функционала (1) соответствует минимуму суммарной энергии растяжения пружин.

При этом решение задачи минимизации функционала (1) полностью аналогично поиску равновесия балки. Зафиксируем ее начальное положение и найдем такое положение пружин, которое отвечает минимуму энергии растяжения. После этого зафиксируем положение концов пружин на балке, освободим балку и дадим ей прийти в механическое равновесие. Далее зафиксируем балку в новом положении и вновь освободим концы пружин.

Эта итерационная процедура сходится – так как на каждом этапе суммарная энергия растяжения пружин уменьшается.

Если же предположить, что балка может упруго отклоняться от прямого вида, то получим следующую картину (рис. 2).

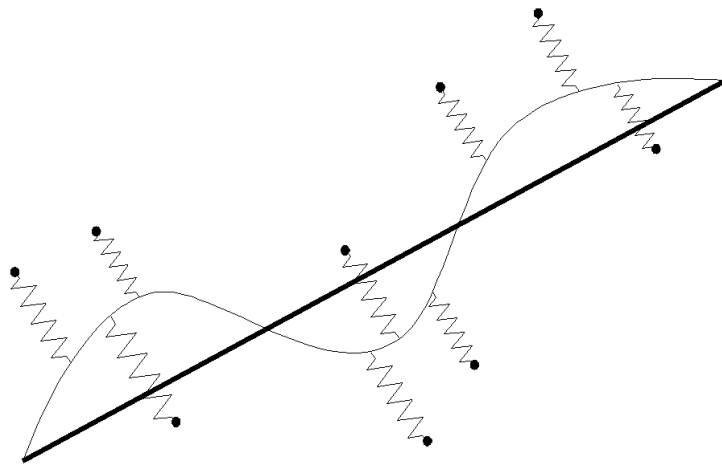


Рис. 2.

Места крепления пружин в балке определяются проекциями на прямую балку (т.к. модель квазилинейная).

При этом возникает задача определения поведения концов балки за границами области данных – описанная выше задача экстраполяции. Формула Карлемана в этом случае имеет аналогию с закреплением бесконечных концов балки на прямой.

Если учесть, что построенная квазилинейная модель является суммой

квазилинейных вектор-функций (3), то можно сделать вывод, что указанный алгоритм допускает нейросетевую интерпретацию.

Действительно, построим такую нейронную сеть, где с каждой кривой из (2) был бы связан один сумматор (в качестве его весов будем использовать координаты вектора y^q), набор из n свободных слагаемых ("порогов") – координат вектора b^q , и n нелинейных преобразователей, каждый из которых вычисляет одну координату точки на кривой по формуле (2).

Действие такого "нейрона" на вектор a входных сигналов специфично: сначала вычисляется проекция данного $t(a)$ на прямую (работает сумматор), далее нелинейные элементы вычисляют $f^q(t(a))$, а затем разность $f_j^q(t(a))$ ($a_j \neq @$) передается следующему нейрону.

При прохождении a по этому конвейеру одновременно накапливается сумма величин $f_j^q(t(a))$ ($a_j = @$). Они и образуют вектор выходных сигналов – предлагаемые значения пропущенных данных. В случае необходимости провести ремонт данных накапливается сумма величин $f_j^q(t(a))$ для каждой координаты j .

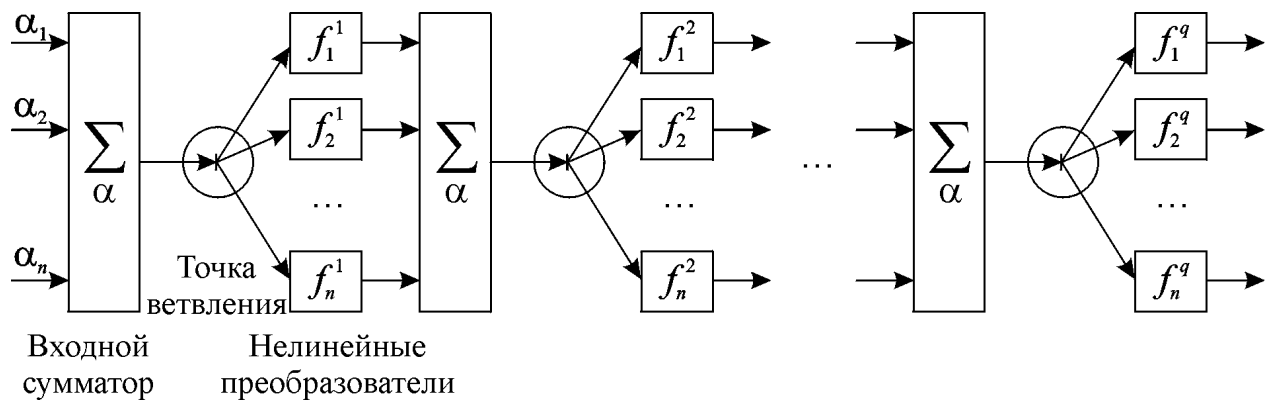


Рис. 3.

В результате получена следующая нейронная сеть (рис. 3). Структура каждого нейрона в ней нестандартна – он имеет один входной сумматор и n нелинейных преобразователей (по размерности вектора данных).

В третьей главе "Самоорганизующиеся многообразия малой размерности" описываются принципы построения и использования самоорганизующихся многообразий малой размерности (Self-Organizing Curve – SOC).

Первоначально вводится понятие главной кривой, как гладкой линии, проходящей через центры масс точек, проецируемых на нее, и рассматривается идея итерационного алгоритма построения главной кривой. Приводится алгоритм Hastie-Stuetzle, который работает итерационно, начиная с главной компоненты множества данных.

Далее в главе рассматривается идея построения самоорганизующихся кривых на основе самоорганизующихся карт Кохонена и описывается алгоритм расчета одномерных нелинейных многообразий.

Пусть SOC определяется набором точек (ядер) $Y = \{y_j\}$ ($j=1..m$), последовательно расположенных на кривой (в первом приближении пусть SOC

– просто ломаная Y) и требуется отобразить на ней набор точек данных $X=\{x_i\}$. Введем преобразование Π , которое каждому вектору $x \in X$ сопоставляет ближайшую к нему точку из Y :

$$x \xrightarrow{\Pi} y_j, \|y_j - x\|^2 \rightarrow \min,$$

каждому ядру y_j сопоставляется его таксон.

$$K_j = \left\{ x \in X \mid x \xrightarrow{\Pi} y_j \right\}.$$

Метод построения SOC напоминает метод динамических ядер, за исключением добавления дополнительных ограничений на связность и нелинейность. Минимизируемая величина строится из следующих слагаемых:

1) мера приближения данных:

$$D_1 = \sum_j \sum_{x \in K_j} \|x - y_j\|^2,$$

2) мера связности (близкие точки на кривой должны переходить в близкие в пространстве данных):

$$D_2 = \sum_j \|y_j - y_{j+1}\|^2,$$

3) мера нелинейности (равномерности):

$$D_3 = \sum_j \|2y_j - y_{j-1} - y_{j+1}\|^2.$$

Таким образом, для построения SOC требуется минимизировать функционал:

$$D = \frac{D_1}{|X|} + \lambda \frac{D_2}{m} + \mu \frac{D_3}{m} \rightarrow \min,$$

где λ, μ – параметры связности и нелинейности – "модули упругости" (деление на число точек $|X|$ и число ядер m означает нормировку "на одно слагаемое" и позволяет для выборок разной мощности использовать одинаковые способы изменения λ и μ).

Аналогично процедуре построения самоорганизующихся кривых строятся соответствующие самоорганизующиеся карты. Приводятся их две основные вариации – с квадратными и гексагональными ячейками сетки.

В отличие от линейных и квазилинейных моделей, при построении самоорганизующихся кривых возникает ряд проблем, основная из которых – попадание в область локального минимума функционала D . Предлагаются возможные пути ее решения – метод "отжига" и многосеточные методы. В первом случае система первоначально ставится в очень жесткие рамки (за счет увеличения коэффициентов в функционале D), которые затем постепенно ослабляются, а во втором в процессе построения кривой изменяется число составляющих ее узлов.

Еще одна проблема заключается в кусочной гладкости полученных многообразий. Поэтому для сглаживания предлагается использовать кусочно-линейные проекторы на ломаную с последующим сглаживанием с

использованием формулы Карлемана, которая при этом еще решает и задачу экстраполяции.

Аналогично линейным и квазилинейным многообразиям, принцип построения самоорганизующихся многообразий малой размерности также имеет ясную механическую интерпретацию.

Искомая упругая балка (кривая – SOC) представляется в виде ломаной, узлы которой свободно соединены с данными (рис. 4).

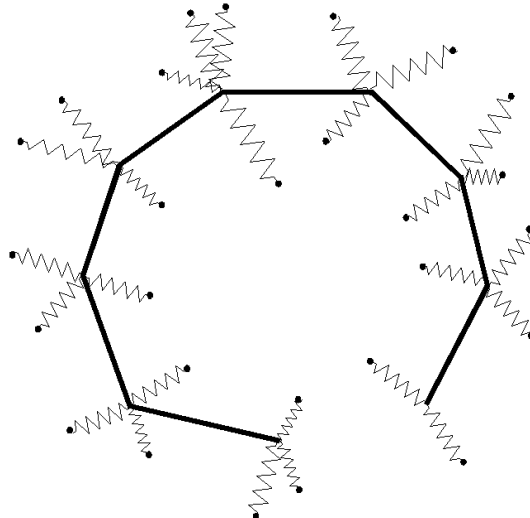


Рис. 4.

Аналогично жесткому случаю, система через несколько итераций придет в равновесие. Причем их число будет конечно, т.к. на каждом шаге суммарная энергия растяжения пружинок уменьшается, а число возможных состояний (способов крепления узлов ломаной пружинками к данным) конечно.

Введенные модули упругости представляют собой соответственно степень притяжения узлов ломаной друг к другу и степень сопротивления изгибу в узлах.

Четвертая глава *"Экспериментальные результаты"* посвящена экспериментальной проверке метода итерационного моделирования неполных данных с помощью многообразий малой размерности.

Совместно с Красноярской государственной медицинской академией был проведен анализ административных территорий Красноярского края по показателям здоровья и здравоохранения – типизация и выделение относительно похожих групп в системе из 49 регионов (административных территорий) Красноярского края с помощью факторного (линейного и нелинейного) и кластерного анализа.

Процесс моделирования данных с пробелами проиллюстрирован также на основе таблицы результатов выборов президентов США, которая содержит 31 предвыборную ситуацию (с 1860 по 1980 гг.), где для каждого выбора в таблице содержатся данные по 12-ти бинарным признакам.

Метод также использовался для верификации связей между двумя динамическими системами – в задачах гелиофизики при анализе различных временных рядов. Из полного ряда годовых значений чисел Вольфа было удалено 50% точек. Для восстановления пропусков использовалась SOC.

Строки исходной таблицы представляли собой m -мерные запаздывающие векторы Такенса, с $m=6$, вида: $a_{kj} = x_k, x_{k+1}, \dots, x_{k+5}$, так что удаление одного отсчета в таблице приводит к удалению соответствующей диагонали. С использованием метода удалось удовлетворительно восстановить даже вершины циклов.

Далее описывается использование метода моделирования неполных данных с помощью многообразий малой размерности при обнаружении факторов, влияющих на течение и прогноз заболевания у больных со сложными нарушениями ритма и проводимости сердца. При этом были замечены различия между разными группами больных и были выявлены факторы, способствующие возникновению ряда осложнений.

В заключении суммированы основные результаты работы и сделаны выводы.

1. Для решения задачи заполнения пропусков и ремонта искаженных данных разработан метод итерационного моделирования неполных данных с помощью многообразий малой размерности. Приведены три вариации метода: начиная с простейших линейных многообразий, продолжая построенными на их основе квазилинейными многообразиями и заканчивая методом главных кривых для данных с пробелами.

2. Для параллельной реализации метода итерационного моделирования данных с пробелами разработан способ построения нейронного конвейера, решающего задачи заполнения пробелов и ремонта данных.

3. Разработаны программные продукты FAMaster и ModelAnalyzer, реализующие предложенные технологии.

4. Численные эксперименты показали высокую эффективность итерационного моделирования неполных данных с помощью многообразий малой размерности. Метод хорошо зарекомендовал себя при решении трудных задач с большим числом пропущенных данных, а в более простых (стандартных) случаях приводит к тем же результатам, что и классические методы статистического анализа.